SDU 2021 summer

Hong Liu

14th December 2021

Abstract

Much of the material here are based on notes from David Galvin and Tim Gowers, and Tao's blog.

Contents

1 Entropy axioms à la Shannon-Khinchin

A central topic of information theory is to efficiently encode complicated sets, say $\mathcal{A}$, using simpler sets, e.g. 0,1-strings. A classical example is sending messages through a (noisy) channel using 0,1-strings. By efficiently, we mean the encoding, as an injection from the complicated sets to simpler ones, is more ideal if it is close to bijection. We shall see that how some of the ideas from information theory can be used in combinatorial problems.

Given a discrete random variable X, we will introduce an information theoretical notion of the *entropy* of X. The entropy of X, denoted by $\mathsf{H}(X)$, is a non-negative real number, measuring the amount of information/surprise/randomness the random variable X carries.

Formally, let X be a random variable taking values from a finite set of alphabets A, write $p_x = \mathsf{Pr}(X = x)$ for each $x \in A$. Then the entropy of X is defined as

$$\mathsf{H}(X) = \sum_{x \in A} p_x \cdot \log \frac{1}{p_x}, \tag{1}$$

where the log is base 2 and we take the convention that $0 \log 0 = 0$.

It is helpful to start with a toy example. Suppose a random variable X is the outcome of a coin flip. What is the entropy of X? If the coin is completely biased, say both of its sides are heads, then the entropy $\mathsf{H}(X) = 1 \log 1 + 0 \log 0 = 0$. This makes intuitive sense: there is no information/surprise from X as we know it will be head. What if the coin is a fair coin with 50/50 chance of landing in head/tail? The entropy would be in this case $\frac{1}{2} \log 2 + \frac{1}{2} \log 2 = 1$. We can also think of entropy as the number of bits of information we gain from the random variable. We shall see that when X is the coin flip outcome, then $\mathsf{H}(X) \leq 1$ with equality if and only if when the coin is a fair one. In general, if X is a random variable over a finite alphabet A of size n, then $\mathsf{H}(X) \leq \log n$ with equality if and only if X is chosen uniformly over A.

Another way of thinking of entropy is to see it as the expected number of bits to specify the random variable. Consider a random variable X uniformly chosen from A and suppose $|A| = 2^k$. Then the entropy of X is $\mathsf{H}(X) = k$. As elements in A are equally likely to appear, we need k bits to specify them. In the above toy example, we have $|A| = 1 = 2^0$ for the biased two-headed coin and $|A| = 2^1$ for the fair coin.

This important property that entropy is maximised when the random variable is uniformly chosen is the key reason why entropy is useful for combinatorial problems. Suppose we want to estimate the cardinality of a set A, then, letting X be a uniformly random element in A, $\mathsf{H}(X) = \log |A|$; thus bounding $|A|$ is the same as bounding the entropy $\mathsf{H}(X)$.

We can also define joint entropy and conditional entropy as follows. Let X, Y be random variables defined over A and B respectively. For each $a \in A$ and $b \in B$, write $p_a = \mathsf{Pr}(X = a)$ and $p_{ab} = \mathsf{Pr}(X = a, Y = b)$, then the *joint entropy* of X and Y is

$$\mathsf{H}(X, Y) = \sum_{a \in A,\ b \in B} p_{ab} \cdot \log \frac{1}{p_{ab}};$$

and the *conditional entropy* of Y given X is

$$\mathsf{H}(Y|X) = \mathsf{E}_{a \in A}\Big(\mathsf{H}(Y|X = a)\Big) = \sum_{a \in A} p_a \cdot \mathsf{H}(Y|X = a).$$

We can think of the conditional entropy $\mathsf{H}(Y|X)$ as the additional information we gain from Y after knowing X.

We will consider only discrete variables taking values from finite sets. We write $x = a \pm b$ for inequalities $a - b \leq x \leq a + b$.

1.1 Shannon-Khinchin entropy axioms

Before being able to use entropy to combinatorial problems, we shall build up some basic theory. Following Gowers's practice, we would like to *forget* the concrete definition of entropy in (1), and do things axiomatically. The advantage is that it saves us from lengthy computation and many arguments are reduced to intuitive calculus.

Below are the Shannon-Khinchin axioms that entropy satisfies:

1. *Invariance.* $\mathsf{H}(X)$ depends only on the probability distribution of X. If $Y = f(X)$ for some function f, then $\mathsf{H}(X, Y) = \mathsf{H}(X)$. If f is bijective, then $\mathsf{H}(Y) = \mathsf{H}(X)$.

2. *Maximality.* If X takes values over A, then $\mathsf{H}(X)$ is maximised when X has the uniform distribution over A.

3. *Extensibility.* If X takes values in A and Y takes values in $B \supseteq A$, and if $\mathsf{Pr}(X = a) = \mathsf{Pr}(Y = a)$ for every $a \in A$, then $\mathsf{H}(Y) = \mathsf{H}(X)$.

4. *Additivity.* $\mathsf{H}(X, Y) = \mathsf{H}(X) + \mathsf{H}(Y|X)$.

5. *Continuity.* $\mathsf{H}(X)$ depends continuously on $\mathsf{Pr}(X = a)$.

6. *Normalisation.* If X is a uniform random variable over two values, then $\mathsf{H}(X) = 1$.

These are very intuitive axioms. Indeed, invariance says that variables carry the same amount of information if they have the same (up to permutation) distribution. Maximality says that the uniform distribution is the most random and the most unpredictable, hence the highest entropy. What about additivity? Well, the information we get from the joint random variable (X, Y) is the same as the information we get from X plus the additional ones we get from Y after revealing X.

We remark that axiom 1–5 determines the entropy up to a constant factor; using axiom 6, we shall see later that the function in (1) is the unique one satisfying all axioms. This is, however, rather unimportant, as we would like to encourage everyone to use these intuitive axioms instead of the dry definition in (1).

1.2 Basic properties of entropy

We shall derive in this section some basic (and intuitive!) properties of entropy using Shannon-Khinchin axioms.

The first property is that if X and Y have nothing to do with each other, then knowing X does not tell us anything about Y.

Lemma 1.1 (Independence)**.** *If X and Y are independent random variables, then*

$$\mathsf{H}(Y|X) = \mathsf{H}(Y) \quad \text{and} \quad \mathsf{H}(X, Y) = \mathsf{H}(X) + \mathsf{H}(Y).$$

In general, if X_i are independent copies of X, then

$$\mathsf{H}(X_1, \ldots, X_n) = n\mathsf{H}(X).$$

Proof. For any x, the distribution of Y given $X = x$ is the same as that of Y as they are independent. Thus, by invariance $\mathsf{H}(Y|X = x) = \mathsf{H}(Y)$, and

$$\mathsf{H}(Y|X) = \sum_x \mathsf{Pr}(X = x)\mathsf{H}(Y|X = x) = \sum_x \mathsf{Pr}(X = x)\mathsf{H}(Y) = \mathsf{H}(Y).$$

Consequently, by additivity, $\mathsf{H}(X, Y) = \mathsf{H}(X) + \mathsf{H}(Y|X) = \mathsf{H}(X) + \mathsf{H}(Y)$. The second assertion follows from induction on n. □

The second one below appeared already in the introduction; recall the biased coin with both sides being head. It says that there is no surprise if there is no uncertainty.

Lemma 1.2. *If X is a random variable taking only one value, then $\mathsf{H}(X) = 0$.*

Proof. Let X_i be independent copies of X. As X takes only one value, by invariance, $\mathsf{H}(X_1, X_2) = \mathsf{H}(X)$. On the other hand, as X_1 and X_2 are independent, by Lemma 1.1 and invariance, $\mathsf{H}(X_1, X_2) = \mathsf{H}(X_1) + \mathsf{H}(X_2) = 2\mathsf{H}(X)$. Then $2\mathsf{H}(X) = \mathsf{H}(X)$ and so $\mathsf{H}(X) = 0$. □

The next one is also intuitive, saying that if a variable takes more values, it is more random, hence higher entropy. In what follows, we denote a random variable X chosen uniformly over a set A as $X \sim A$.

Lemma 1.3 (Monotonicity 1)**.** *Let $X \sim A$ and $Y \sim B$. If $A \subseteq B$, then $\mathsf{H}(X) \le \mathsf{H}(Y)$ with equality if and only if $A = B$.*

Proof. By extensibility, we can think of X as taking values in B (i.e. $\mathsf{Pr}(X = a) = 1/|A|$ for each $a \in A$ and $\mathsf{Pr}(X = b) = 0$ for each $b \in B \setminus A$). Then $\mathsf{H}(X) \le \mathsf{H}(Y)$ by maximality. If $A = B$, then invariance implies the equality $\mathsf{H}(X) = \mathsf{H}(Y)$.

To see the strict inequality, suppose $|A| < |B|$. If $|A| = 1$, then $\mathsf{H}(X) = 0$ due to Lemma 1.2. On the other hand, let $B' \subseteq B$ be a subset of size 2 and let $Z \sim B'$, then $\mathsf{H}(Y) \ge \mathsf{H}(Z)$. Thus, by normalisation, $\mathsf{H}(Y) \ge \mathsf{H}(Z) = 1 > 0 = \mathsf{H}(X)$.

Suppose then $|A| \ge 2$, and so $\mathsf{H}(X) \ge 1$. Let X_i, Y_i be independent copies of X and Y respectively. Choose n sufficiently large so that $|A|^n \le |B|^{n-1}$, then by Lemma 1.1,

$$n\mathsf{H}(X) = \mathsf{H}(X_1, \dots, X_n) \le \mathsf{H}(Y_1, \dots, Y_{n-1}) = (n-1)\mathsf{H}(Y).$$

As $\mathsf{H}(X) \ge 1$, we get the strict inequality $\mathsf{H}(X) < \mathsf{H}(Y)$. □

For a random variable with rational probabilities on atoms, it is helpful to link it to a uniform distribution as follows.

Construction 1.4. Let X be a random variable taking values in A such that for each $a \in A$, $\mathsf{Pr}(X = a) = \frac{m_a}{n}$ for some $m_a \in \mathbb{N}$. Let $U \sim [n]$. We can think of X being determined by U as follows. Let V_a, $a \in A$, be a partition of $[n]$ with $|V_a| = m_a$. Let X' be the random variable over A such that $X' = a$ if $U \in V_a$. Then X, X' are identically distributed, and by invariance $\mathsf{H}(X) = \mathsf{H}(X')$. Furthermore, as we define X' from U, by invariance, $\mathsf{H}(X', U) = \mathsf{H}(U)$. Another useful fact here is that condition on $X' = a$, Y is uniformly distributed over V_a, i.e. the random variable $(U|X' = a) \sim V_a$.

We can also think of U in terms of X. Let U' be the random variable over $[n]$ such that if $X = a$ then U' is uniform over V_a.[1] Then $U, U' \sim [n]$ are identically distributed.

Entropy assigns non-negative values to discrete random variables. This is obvious from the definition in (1). Let us prove it using the axioms.

Lemma 1.5. *Let X be a random variable taking values in a finite set A, then $\mathsf{H}(X) \ge 0$.*

Proof. Suppose X takes values in A with rational probabilities. Let $U \sim [n]$ and X' be as in Construction 1.4, so $\mathsf{H}(X') = \mathsf{H}(X)$ and $\mathsf{H}(X', U) = \mathsf{H}(U)$. By additivity, $\mathsf{H}(X', U) = \mathsf{H}(X') + \mathsf{H}(U|X')$. Thus,

$$\mathsf{H}(X) = \mathsf{H}(X') = \mathsf{H}(X', U) - \mathsf{H}(U|X') = \mathsf{H}(U) - \mathsf{H}(U|X'),$$

and it suffices to show $\mathsf{H}(U|X') \le \mathsf{H}(U)$. Recall that for each $a \in A$, $(U|X' = a) \sim V_a \subseteq [n]$ and $U \sim [n]$. Thus, Lemma 1.3 entails that $\mathsf{H}(U|X' = a) \le \mathsf{H}(U)$, and so

$$\mathsf{H}(U|X') = \sum_{a \in A} \mathsf{Pr}(X' = a)\mathsf{H}(U|X' = a) \le \sum_{a \in A} \mathsf{Pr}(X' = a)\mathsf{H}(U) = \mathsf{H}(U).$$

The general case follows from the above case and continuity axiom. Indeed, as the set of rationals is dense in reals, we can approximate the atom probability $\mathsf{Pr}(X = a)$ arbitrarily closely by multiples of $1/n$ for large enough n. □

[1] Note that X does *not* determine U'.

The following property says that if Y is determined by X, then it carries no more information than X.

Lemma 1.6 (Monotonicity 2)**.** *Given random variables X, Y, if $Y = f(X)$ for some function f, then $\mathsf{H}(Y) \leq \mathsf{H}(X)$, with equality if f is bijective.*

Proof. The equality case when f is bijective is simply invariance. For the inequality, as Y is determined by X, by invariance, $\mathsf{H}(X, Y) = \mathsf{H}(X)$. On the other hand, by additivity and that entropy is non-negative (Lemma 1.5),

$$\mathsf{H}(X) = \mathsf{H}(X, Y) = \mathsf{H}(Y) + \mathsf{H}(X|Y) \geq \mathsf{H}(Y)$$

as desired. □

When the outcome is not absolute certainty, then entropy is positive.

Lemma 1.7. *Let X be a random variable taking at least two values with positive probability, then $\mathsf{H}(X) > 0$.*

Proof. Let A be the set over which X is defined, and set $c = \max_{a \in A} \mathsf{Pr}(X = a)$. By assumption, $c < 1$. Thus, for any $\varepsilon > 0$, we can choose n large enough so that $c^n < \varepsilon$. Letting $X_1, \ldots, X_n$ be independent copies of X, we see that $(X_1, \ldots, X_n)$ takes any value in A^n with probability at most $c^n < \varepsilon$. So we can partition A^n into two sets A_0, A_1, each with probability $\frac{1}{2} \pm \varepsilon$. Define random variable Y such that $Y = i$ if $(X_1, \ldots, X_n) \in A_i$, $i \in \{0, 1\}$.

Now, by normalisation and continuity, $\mathsf{H}(Y) > 0$. By Lemma 1.1, $\mathsf{H}(X_1, \ldots, X_n) = n\mathsf{H}(X)$. On the other hand, as Y is determined by $(X_1, \ldots, X_n)$, by Lemma 1.6, $\mathsf{H}(X_1, \ldots, X_n) \geq \mathsf{H}(Y)$. So $\mathsf{H}(X) \geq \frac{1}{n}\mathsf{H}(Y) > 0$ as desired. □

The last one below follows immediately from additivity and induction.

Lemma 1.8 (Chain rule)**.** *Let $X_1, \ldots, X_n$ be random variables. Then*

$$\mathsf{H}(X_1, \ldots, X_n) = \mathsf{H}(X_1) + \mathsf{H}(X_2|X_1) + \cdots + \mathsf{H}(X_n|X_1, \ldots, X_{n-1}) = \sum_{i \in [n]} \mathsf{H}(X_i|X_1, \ldots, X_{i-1}).$$

1.3 Subadditivity and Shearer's lemma

We shall see in this section yet another key property of entropy, subadditivity and its generalisation, Shearer's lemma.

Let us start with an intuitive statement, stating that dropping conditioning can only increase entropy. This makes sense as knowing Y would only decrease the amount of information we get from X.

Lemma 1.9 (Dropping conditioning)**.** *Let X, Y, Z be random variables. Then*

$$\mathsf{H}(Y|X) \leq \mathsf{H}(Y) \quad \textit{and} \quad \mathsf{H}(Z|Y, X) \leq \mathsf{H}(Z|Y).$$

Proof. Let us first prove the special case when X is uniformly distributed. In this case, by maximality, $\mathsf{H}(X|Y = b) \leq \mathsf{H}(X)$ for any b, and so $\mathsf{H}(X|Y) \leq \mathsf{H}(X)$. Consequently, by additivity,

$$\mathsf{H}(Y|X) = \mathsf{H}(X, Y) - \mathsf{H}(X) \leq \mathsf{H}(X, Y) - \mathsf{H}(X|Y) = \mathsf{H}(Y).$$

Next, by continuity, it suffices to consider the case when X takes values with rational probabilities. Let $U' \sim [n]$ be as in Construction 1.4, so $(U'|X = a) \sim V_a$. As for any a, $(U'|X = a)$ is uniformly distributed, the special case above implies

$$\mathsf{H}\Big(Y \big| (U'|X = a)\Big) \leq \mathsf{H}(Y),$$

and so $\mathsf{H}(Y|U', X) \le \mathsf{H}(Y)$. On the other hand, $(U'|X = a) \sim V_a$ implies that given $X = a$, U' and Y are conditionally independent. Then Lemma 1.1 entails

$$\mathsf{H}(Y|U', X = a) = \mathsf{H}(Y|X = a).$$

Hence $\mathsf{H}(Y|X) = \mathsf{H}(Y|U', X) \le \mathsf{H}(Y)$ as desired.

The second statement follows from the first one. □

We can now prove that entropy is *subadditive*. It states that the information we get jointly from $X_1, \dots, X_n$ is no more than the combine of the information we get from each of the X_is.

Lemma 1.10 (Subadditivity)**.** *Let $X_1, \dots, X_n$ be random variables. Then*

$$\mathsf{H}(X_1, \dots, X_n) \le \sum_{i \in [n]} \mathsf{H}(X_i).$$

Proof. As conditioning less can only increase the entropy (Lemma 1.9), together with additivity, $\mathsf{H}(X_1, X_2) = \mathsf{H}(X_1) + \mathsf{H}(X_2|X_1) \le \mathsf{H}(X_1) + \mathsf{H}(X_2)$. The conclusion then follows from induction on n. □

A generalisation of subadditivity was given by Shearer.

Lemma 1.11 (Shearer's lemma)**.** *Let $\mathcal{F}$ be a family of subsets of $[n]$ (possibly with repeats) such that each coordinate $i \in [n]$ is contained in at least k members of $\mathcal{F}$. Then for a random vector $(X_1, \dots, X_n)$,*

$$\mathsf{H}(X_1, \dots, X_n) \le \frac{1}{k} \sum_{F \in \mathcal{F}} \mathsf{H}(X_F),$$

where X_F is the vector $(X_i : i \in F)$.

Subadditivity is the special case of Shearer's lemma in which $\mathcal{F}$ consists of all singletons in $[n]$ and $k = 1$.

We shall give a proof of the following equivalent probabilistic version. It states that the entropy of a random vector can be bounded in terms of the expected entropy of a random projection; and the bound is more effective if the random projection covers every coordinate with decent probability.

Lemma 1.12 (Shearer's lemma, probabilistic version)**.** *Let F be a random subset of $[n]$ such that for each coordinate $i \in [n]$, $\mathsf{Pr}(i \in F) \ge \mu$. Then for a random vector $(X_1, \dots, X_n)$,*

$$\mathsf{H}(X_1, \dots, X_n) \le \frac{1}{\mu} \mathsf{E}_F \mathsf{H}(X_F),$$

where X_F is the vector $(X_i : i \in F)$.

Proof. Order the random elements in F as $i_1 < \dots < i_k$. Then by chain rule (Lemma 1.8),

$$\mathsf{H}(X_F) = \mathsf{H}(X_{i_1}) + \mathsf{H}(X_{i_2}|X_{i_1}) + \dots + \mathsf{H}(X_{i_k}|X_{i_1}, \dots, X_{i_{k-1}}).$$

Write $\mathsf{H}(X_i|X_{<i})$ for $\mathsf{H}(X_i|X_1, \dots, X_{i-1})$. Conditioning more (which only reduces entropy by Lemma 1.9) and taking expectation, we get

$$\begin{aligned}
\mathsf{E}_F \mathsf{H}(X_F) &\ge \mathsf{E}_F \sum_{i \in F} \mathsf{H}(X_i|X_{<i}) \\
&= \sum_{i \in [n]} \mathsf{Pr}(i \in F) \cdot \mathsf{H}(X_i|X_{<i}) \\
&\ge \mu \sum_{i \in [n]} \mathsf{H}(X_i|X_{<i}) \\
&= \mu \mathsf{H}(X_1, \dots, X_n),
\end{aligned}$$

where the last equality follows from chain rule. □

1.4 Axioms determine entropy function uniquely

We wrap up the theory developing by proving that Shannon-Khinchin axioms determine the entropy function uniquely.

Let us first see that the function above satisfies all axioms. Let us show maximality and additivity; the others are trivial.

Recall Jensen's inequality. If f is concave below, the inequality goes the other way.

Lemma 1.13. *Let X be a random variable and f be a convex function, then*

$$f(\mathsf{E}X) \leq \mathsf{E}(f(X)).$$

Maximality follows from Jensen's inequality.

Proposition 1.14. *Let X be a random variable defined over A and let $\mathsf{H}(X) = \sum_{x\in A} p_x \cdot \log \frac{1}{p_x}$, where $p_x = \mathsf{Pr}(X = x)$. Then $\mathsf{H}(X)$ satisfies maximality.*

Proof. As $f(x) = \log x$ is concave, by Jensen's inequality,

$$\mathsf{H}(X) = \sum_{x\in A} p_x \cdot \log \frac{1}{p_x} \leq \log \Big(\sum_{x\in A} p_x \cdot \frac{1}{p_x} \Big) = \log |A|,$$

with equality when all p_x are equal, i.e. when X is uniform over A. □

Proposition 1.15. *Let X be a random variable defined over A and let $\mathsf{H}(X) = \sum_{a\in A} p_a \cdot \log \frac{1}{p_a}$, where $p_a = \mathsf{Pr}(X = a)$. Then $\mathsf{H}(X)$ satisfies additivity.*

Proof. Let Y be a random variable defined over B. Write $q_b = \mathsf{Pr}(Y = b)$, $p_{ab} = \mathsf{Pr}(X = a, Y = b)$ and $q_{b|a} = \mathsf{Pr}(Y = b|X = a)$. We want to show $\mathsf{H}(X,Y) = \mathsf{H}(X) + \mathsf{H}(Y|X)$.

Firstly, as

$$p_{ab} = \mathsf{Pr}(X = a, Y = b) = \mathsf{Pr}(X = a) \cdot \mathsf{Pr}(Y = b|X = a) = p_a q_{b|a},$$

we can rewrite the joint entropy:

$$\begin{aligned} \mathsf{H}(X,Y) &= \sum_{a\in A,\ b\in B} p_{ab} \log \frac{1}{p_{ab}} \\ &= \sum_{a\in A,\ b\in B} p_{ab} \Big(\log \frac{1}{p_a} + \log \frac{1}{q_{b|a}} \Big). \end{aligned}$$

Note that $\sum_{b\in B} p_{ab} = p_a$, so the first term in the sum is

$$\sum_{a\in A,\ b\in B} p_{ab} \log \frac{1}{p_a} = \sum_{a\in A} \log \frac{1}{p_a} \big(\sum_{b\in B} p_{ab} \big) = \sum_{a\in A} p_a \log \frac{1}{p_a} = \mathsf{H}(X).$$

We are left to show the second term in the sum is $\mathsf{H}(Y|X)$. Indeed,

$$\begin{aligned} \sum_{a\in A,\ b\in B} p_{ab} \log \frac{1}{q_{b|a}} &= \sum_{a\in A} p_a \sum_{b\in B} q_{b|a} \log \frac{1}{q_{b|a}} \\ &= \sum_{a\in A} p_a \mathsf{H}(Y|X = a) \\ &= \mathsf{H}(Y|X), \end{aligned}$$

as desired. □

We now show that the choice of entropy function in (1) is the unique one satisfying all the axioms. First, we need the following.

Lemma 1.16. *If X is a random variable uniformly distributed over A, then* $\mathsf{H}(X) = \log |A|$.

Proof. Let $|A| = n$ and consider the special case that $n = 2^k$. Let Y be uniformly chosen from $[2]$, then $\mathsf{H}(Y) = 1$ by normalisation. Let $Y_1, \dots, Y_k$ be indepdendent copies of Y, then by Lemma 1.1, $\mathsf{H}(Y_1, \dots, Y_k) = k\mathsf{H}(Y) = k$. As $(Y_1, \dots, Y_k)$ is uniform over $[2]^k$, by invariance, $\mathsf{H}(X) = \mathsf{H}(Y_1, \dots, Y_k) = k$.

Now, for general n, let $\delta = \mathsf{H}(X) - \log n$. Take independent copies $X_1, \dots, X_r$ of X, then again by Lemma 1.1, $\mathsf{H}(X_1, \dots, X_r) = r\mathsf{H}(X)$. So if $2^k \leq n^r \leq 2^{k+1}$, we have from Lemma 1.3 and the special case above that $k \leq r\mathsf{H}(X) \leq k + 1$, or $\frac{k}{r} - \log n \leq \delta \leq \frac{k+1}{r} - \log n$. As $\frac{k}{r} - \log n \leq 0$ and $\frac{k+1}{r} - \log n \geq 0$, we get that $|\delta| \leq \frac{1}{r}$. It follows that $\delta = 0$ as the choice of r is arbitrary. Thus, $\mathsf{H}(X) = \log n$. □

Lemma 1.17. *Let X be a random variable defined over A. If $\mathsf{H}(X)$ satisfies all six Shannon-Khinchin axioms for entropy, then $\mathsf{H}(X) = \sum_{a \in A} p_a \cdot \log \frac{1}{p_a}$, where $p_a = \mathsf{Pr}(X = a)$.*

Proof. By continuity, we may assume that there is some $n \in \mathbb{N}$ such that $p_a = \frac{m_a}{n}$ for some $m_a \in \mathbb{N}$ for each $a \in A$. Let $U \sim [n]$, X' be as in Construction 1.4, so $\mathsf{H}(X) = \mathsf{H}(X')$, $\mathsf{H}(X', U) = \mathsf{H}(U)$ and $(U|X' = a) \sim V_a$.

As $(U|X' = a) \sim V_a$, where $|V_a| = p_a n$, and $U \sim [n]$, by Lemma 1.16 and that X, X' are identically distributed, we get that $\mathsf{H}(U) = \log n$ and that

$$\mathsf{H}(U|X') = \sum_{a \in A} \mathsf{Pr}(X' = a)\mathsf{H}(U|X' = a) = \sum_{a \in A} p_a \log(p_a n).$$

Then, by additivity,

$$\begin{aligned} \mathsf{H}(X) = \mathsf{H}(X') &= \mathsf{H}(X', U) - \mathsf{H}(U|X') \\ &= \mathsf{H}(U) - \sum_{a \in A} p_a \log(p_a n) \\ &= \log n - \sum_{a \in A} p_a(\log p_a + \log n) \\ &= \sum_{a \in A} p_a \log \frac{1}{p_a}, \end{aligned}$$

as desired. □

1.5 Conditional verions of basic properties

The basic properties we have shown so far for entropy can be easily extended to conditional entropy. We will summarise in this section these properties. Below, X, Y, Z, X_is are discrete random variables and E is some event, and f is a deterministic function. We write $\text{range}(X|E)$ for the set of values X takes with positive probability conditioning on E.

- *Maximality.* $\mathsf{H}(X|E) \leq \log |\text{range}(X|E)|$ with equality if and only if $X|E \sim \text{range}(X|E)$. This corresponds to the intuition that under uniform measure, no choice is favourable and so the outcome is the hardest to predict, hence carrying the most information.
- *Chain rule.* $\mathsf{H}(X_1, \dots, X_n|Y) = \sum_{i \in [n]} \mathsf{H}(X_i|X_1, \dots, X_{i-1}, Y)$.
- *Monotonicity.* If $X = f(Y)$, then $\mathsf{H}(X) \leq \mathsf{H}(Y)$. More generally, if $X = f(Y, Z)$, then $\mathsf{H}(X|Z) \leq \mathsf{H}(Y|Z)$. If $Y = f(Z)$, then $\mathsf{H}(X|Z) \leq \mathsf{H}(X|Y)$. The last one means that if we reveal less upfront (Y instead of Z), then we get more information later.

- *Dropping conditioning.* $\mathsf{H}(Z|Y,X) \le \mathsf{H}(Z|Y)$.
- *Subadditivity.* $\mathsf{H}(X_1,\dots,X_n|Y) \le \sum_{i\in[n]} \mathsf{H}(X_i|Y)$ with equality if X_is are conditionally independent given Y.
- *Shearer's lemma.* Given random $F \subseteq [n]$ with $\mathsf{Pr}(i \in F) \ge \mu$ for each $i \in [n]$, then
$$\mathsf{H}(X_1,\dots,X_n|Y) \le \frac{1}{\mu}\mathsf{E}_F\mathsf{H}(X_F|Y).$$

There is another useful inequality. We will state it without proof.

- *Gibbs inequality.* If X, Y take values in the same finite set A, writing $p_a = \mathsf{Pr}(X = a)$ and $q_a = \mathsf{Pr}(Y = a)$, then
$$\mathsf{H}(X) = \sum_{a\in A} p_a \log\frac{1}{p_a} \le \sum_{a\in A} p_a \log\frac{1}{q_a},$$
with equality if and only if $p_a = q_a$ for all $a \in A$.

This is the same as saying that the *Kullback–Leibler divergence* or *relative entropy*, denoted by $D_{\mathrm{KL}}(X\|Y)$, is non-negative:
$$D_{\mathrm{KL}}(X\|Y) = \sum_{a\in A} p_a \log\frac{p_a}{q_a} \ge 0.$$

2 Applications of entropy

2.1 Volume of Hamming balls

The first application of entropy is to estimate the volume of Hamming balls. We need a piece of notation.

Definition 2.1. The *binary entropy function* $h : [0,1] \to \mathbb{R}$ is defined as
$$h(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}.$$

Note that $h(p)$ is the entropy of a Bernoulli random variable with probability p, i.e. $h(p) = \mathsf{H}(X)$, where $X \sim \mathrm{Binom}(1,p)$. It is easy to show that $h(p)$ is increasing from 0 to 1 when $p \in [0, \frac{1}{2}]$ and decreasing down to 0 when $p \in [\frac{1}{2}, 1]$.

We can write binomial coefficients using binary entropy function. Using Stirling's formula that $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$ as $n \to \infty$, a simple calculation shows that for any $p \in (0,1)$,
$$\binom{n}{pn} \sim \frac{2^{h(p)n}}{\sqrt{2\pi np(1-p)}}. \tag{2}$$

The *Hamming distance* between two vectors is the number of coordinates on which they differ. Given a binary vector $v \in \{0,1\}^n$, the *Hamming ball* of radius r around it is the set of all vectors with Hamming distance at most r from v. By *volume* of a Hamming ball, we mean its size. Note that the volume of a Hamming ball of radius r is precisely $\sum_{i\le r}\binom{n}{i}$.

We can estimate the volume of a Hamming ball using binary entropy function, which is fairly tight considering the asymptotics in (2). The bound below can of course be derived again using Stirling's formula. The point is that using entropy, we can have a calculation-free proof. All we need below is maximality and subadditivity.

Theorem 2.2. *Let $0 < p \le 1/2$. Then for all n,*

$$\sum_{i \le pn} \binom{n}{i} \le 2^{h(p)n}.$$

Proof. Let A be the set of all subsets of $[n]$ of size at most pn, then we want to show that $|A| = \sum_{i \le pn} \binom{n}{i} \le 2^{h(p)n}$. Let $X \sim A$, then $\mathsf{H}(X) = \log |A|$ (Lemma 1.16). It then suffices to show $\mathsf{H}(X) \le h(p)n$.

Now, think of X as the random vector $(X_1, \ldots, X_n)$, where X_i is the indicator function for $\{i \in X\}$. Then by subadditivity $\mathsf{H}(X) \le \sum_{i \in [n]} \mathsf{H}(X_i)$. Note that $\mathsf{H}(X_i) = h(\alpha_i)$, where $\alpha_i = \mathsf{Pr}(i \in X)$. As X has size at most pn, $\alpha_i \le p \le \frac{1}{2}$ for each $i \in [n]$. As $h(x)$ is increasing when $x \le \frac{1}{2}$, it follows that $\mathsf{H}(X) \le \sum_{i \in [n]} h(\alpha_i) \le h(p)n$ as desired. □

We can use the above estimate to get the following weak form of the Chernoff concentration bound. We leave its proof as exercise.

Corollary 2.3. *Let X be a binomial random variable $X \sim$ Binom$(n, \frac{1}{2})$ with standard deviation $\sigma = \frac{\sqrt{n}}{2}$. Then for any $c \ge 0$,*

$$\mathsf{Pr}\Big(\big|X - \frac{n}{2}\big| \ge c\sigma\Big) \le 2^{-\frac{c^2}{2}+1}.$$

2.2 Loomis-Whitney inequality

Loomis-Whitney inequality in geometry estimates the volume of a n-dimensional body by the volumes of its $(n-1)$-dimensional projections.

For a measurable body K in $\mathbb{R}^n$, we write vol(B) for its volume. For each $i \in [n]$, let $\pi_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$ be the projection to the hyperplane $x_i = 0$, that is,

$$\pi_i(x_1, \ldots, x_n) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

Theorem 2.4 (Loomis-Whitney inequality)**.** *Let K be a measurable body in $\mathbb{R}^n$. Then*

$$\mathrm{vol}(K) \le \prod_{i \in [n]} \mathrm{vol}\big(\pi_i(K)\big)^{\frac{1}{n-1}}.$$

This inequality is tight: equality holds when K is an axis-aligned box.

Proof. By standard scaling and limiting argument, we may assume that K is a union of axis-aligned unit cubes. Let X be a uniform random cube in K, then $\mathsf{H}(X) = \log \mathrm{vol}(K)$ (Lemma 1.16). As each cube is uniquely determined by its center, we can identify it with the center. So $\mathsf{H}(X) = \mathsf{H}(X_1, \ldots, X_n)$, where X_i is the i-th coordinate of X's center. It thus suffices to show that

$$\mathsf{H}(X_1, \ldots, X_n) \le \frac{1}{n-1} \sum_{i \in [n]} \log \mathrm{vol}\big(\pi_i(K)\big).$$

Let $\pi \sim \{\pi_1, \ldots, \pi_n\}$ be a uniform random $(n-1)$-dimensional projection, and let C be the random subset of $[n]$ recording the set of coordinates π projects to. That is, for any vector $(Z_1, \ldots, Z_n) \in \mathbb{R}^n$, $\pi(Z) = Z_C := (Z_i : i \in C)$. As π is uniformly chosen, $\mathsf{Pr}(i \in C) = \mu = \frac{n-1}{n}$.

Then by Shearer's lemma (Lemma 1.12),

$$\begin{aligned}\mathsf{H}(X_1,\ldots,X_n) &\le \frac{n}{n-1}\mathsf{E}_C\mathsf{H}(X_C)\\ &= \frac{n}{n-1}\mathsf{E}_\pi\mathsf{H}(\pi(X))\\ &= \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i\in[n]}\mathsf{H}(\pi_i(X))\\ &\le \frac{1}{n-1}\sum_{i\in[n]}\log\operatorname{vol}\big(\pi_i(K)\big),\end{aligned}$$

where the last inequality follows from maximality and that $\pi_i(X)$ takes values in $\pi_i(K)$. □

2.3 Triangle maximisation

A typical application of Kruskal-Katona theorem is the triangle maximisation problem: if a graph has m edges, how many triangles it can have? Here we show how to use Shearer's lemma to get the same bound (again without much calculation!). Intuitively, given the number of edges, the number of triangles is maximised when edges are clustered together. Say $m = \binom{k}{2}$, then we can pack edges into a clique on k vertices, in which case we have $\binom{k}{3} \sim \frac{(2m)^{3/2}}{6}$ triangles. We will prove this asymptotically tight bound.

Theorem 2.5. *An m-edge graph has at most $\frac{(2m)^{3/2}}{6}$ triangles.*

Proof. Let T be the set of labeled triangles in G and let $t = |T|$. Take $X = (X_1, X_2, X_3) \sim T$, where X_is are vertices of X, then $\mathsf{H}(X) = \log t$. As there are $3! = 6$ ways to label a triangle, it suffices to show that $\mathsf{H}(X) \le \frac{3}{2}\log(2m)$.

Let $F \sim \binom{[3]}{2}$, then $\mathsf{Pr}(i \in F) = \mu = \frac{2}{3}$. Thus, by Shearer's lemma (Lemma 1.12),

$$\mathsf{H}(X) \le \frac{3}{2}\ \mathsf{E}_F\mathsf{H}(X_F) \le \frac{3}{2}\log(2m),$$

where the last inequality follows from maximality and the fact that X_F takes values on the set of all labeled edges. □

2.4 Triangle-intersecting family

A family of graphs on vertex set $[n]$ is *triangle-intersecting* if every two graphs in the family has a triangle in common. Ellis, Filmus and Friedgut showed that a triangle-intersecting family has size at most $2^{\binom{n}{2}-3}$. This bound is tight. As in Erdős-Ko-Rado, we can take the family of all graphs containing a fixed triangle, which has size $2^{\binom{n}{2}-3}$. Here we prove a weaker bound.

Theorem 2.6. *A triangle-intersecting family $\mathcal{F}$ on $[n]$ has size at most $2^{\binom{n}{2}-2}$.*

Proof. Let $X \sim \mathcal{F}$, then we need to show $\mathsf{H}(X) = \log|\mathcal{F}| \le \binom{n}{2} - 2$. Think of X as a random vector $(X_e : e \in \binom{[n]}{2})$, where X_e is the indicator function of $\{e \in X\}$.

Consider now a uniform random set $R \subseteq [n]$ and let G be the random graph consisting of two cliques, one on R and the other on $[n] \setminus R$. Then $\mathsf{E}_G(e(G)) = \frac{1}{2}\binom{n}{2}$, and for each $e \in \binom{[n]}{2}$, $\mathsf{Pr}(e \in G) = \mu = \frac{1}{2}$. Shearer's lemma then implies

$$\mathsf{H}(X) \le 2\ \mathsf{E}_G\mathsf{H}(X_G).$$

The clever idea here is that the family $\mathcal{F}$ restricted to G is an intersecting family: any pair of graphs in $\mathcal{F}$ share a triangle, which must intersect G. Thus, X_G takes values in an intersecting family on edge set of G, which has size at most $2^{e(G)-1}$. So maximality implies $\mathsf{H}(X_G) \le e(G)-1$, and we get that $\mathsf{H}(X) \le 2\mathsf{E}_G(e(G)-1) = \binom{n}{2} - 2$ as desired. □

2.5 Brégman's theorem

The next application of entropy is a proof of Brégman's theorem due to Radhakrishnan on the maximum permanent of a 0,1-matrix with given row sums.

The *permanent* of an $n \times n$ matrix A is the sum

$$\operatorname{perm}(A) = \sum_{\sigma \in S_n} \prod_{i \in [n]} A_{i\sigma(i)},$$

where S_n is the symmetric group on $[n]$. The permanent is like the determinant but without the signs. This makes all the difference: while there is efficiently algorithm to compute determinant, it is computationally difficult to determine permanent.

An equivalent way of looking at permanent of a $0,1$-matrix A is to see it as the number of perfect matchings in a bipartite graph G with partite sets U and V, each of size n. Indeed, we can view A as a bipartite adjacency matrix of G: rows and columns represent vertices in U and V respectively. Then each permutation $\sigma \in S_n$ that contributes 1 to $\operatorname{perm}(A)$ corresponds to a perfect matching in G. Note that this correspondence is bijective.

Brégman's theorem offers an upper bound on the permanent of $0,1$-matrix.

Theorem 2.7 (Brégman's theorem)**.** *Let A be an $n \times n$ $0,1$-matrix with row sums $d_1, \dots, d_n$. Then*

$$\operatorname{perm}(A) \le \prod_{i \in [n]} (d_i!)^{\frac{1}{d_i}}.$$

When viewing as perfect matchings in bipartite graphs, the row sums are the degrees of vertices in one partite set. Brégman's theorem states that given the degree sequence $(d_1, \dots, d_n)$ on one side of the bipartite graph, then the number of perfect matchings is at most $\prod_{i \in [n]} (d_i!)^{\frac{1}{d_i}}$. Note that this upper bound is tight: assume $d|n$ and consider the n-vertex graph that is a union of $\frac{n}{d}$ disjoint copies of $K_{d,d}$.

We will present Radhakrishnan's proof for the graph version.

Proof. View A as the bipartite adjacency matrix of a bipartite graph G with partite sets U and V, and so the degree sequence of U is $(d_1, \dots, d_n)$ and $\operatorname{perm}(A) = |M|$, where M is the set of perfect matchings in G. As usual, take $\sigma \sim M$, then we need to show

$$\mathsf{H}(\sigma) = \log |M| \le \sum_{i \in [n]} \frac{\log(d_i!)}{d_i}.$$

Fix an ordering $\tau : v_1, \dots, v_n$ of U, think of σ as the random vector $(\sigma(v_1), \dots, \sigma(v_n))$, where $\sigma(v_i) \in V$ is the other endpoint of the edge in σ containing v_i. Then by chain rule,

$$\mathsf{H}(\sigma) = \mathsf{H}(\sigma(v_1)) + \mathsf{H}(\sigma(v_2)|\sigma(v_1)) + \cdots + \mathsf{H}(\sigma(v_n)|\sigma(v_1), \dots, \sigma(v_{n-1})).$$

Fix v_k, then after revealing $\sigma(v_1), \dots, \sigma(v_{k-1})$, $\sigma(v_k)$ has to take values in the set of neighbours of v_k that are not equal to any of $\sigma(v_1), \dots, \sigma(v_{k-1})$. Let $d_{\tau,k-1}(v_k)$ be the number of such neighbours, then maximality infers that $\mathsf{H}(\sigma(v_k)|\sigma(v_1), \dots, \sigma(v_{k-1})) \le \mathsf{E}_\sigma \log(d_{\tau,k-1}(v_k))$ and so

$$\mathsf{H}(\sigma) \le \mathsf{E}_\sigma \sum_{k \in [n]} \log(d_{\tau,k-1}(v_k)).$$

The problem now is that we know nothing about how many neighbours of v_k have been used before $\sigma(v_k)$. Radhakrishnan's idea is to take a random ordering τ, then we can know how $d_{\tau,k-1}(v_k)$ behaves in expectation.

Now, fix a choice for σ in M and let $\tau : v_1, \dots, v_n$ be a uniform random ordering of U. Fix a vertex $v \in U$ and consider its contribution to the sum $\sum_{k \in [n]} \log(d_{\tau,k-1}(v_k))$. As τ is a uniform random ordering, the number of used neighbours of v before $\sigma(v)$ is equally likely to be any of $0, \dots, d(v) - 1$. That is, if $v = v_k$, then $d_{\tau,k-1}(v_k) \sim [d(v_k)]$. Thus,

$$\mathsf{E}_\tau \sum_{k \in [n]} \log(d_{\tau,k-1}(v_k)) = \sum_{v \in U} \frac{1}{d(v)} \sum_{i \in [d(v)]} \log i = \sum_{i \in [n]} \frac{\log(d_i!)}{d_i}.$$

Thus, averaging over both τ and σ, we have

$$\mathsf{E}_\tau \mathsf{E}_\sigma \sum_{k \in [n]} \log(d_{\tau,k-1}(v_k)) = \sum_{i \in [n]} \frac{\log(d_i!)}{d_i},$$

implying that there is a choice of τ such that

$$\mathsf{H}(\sigma) \le \mathsf{E}_\sigma \sum_{k \in [n]} \log(d_{\tau,k-1}(v_k)) \le \sum_{i \in [n]} \frac{\log(d_i!)}{d_i}$$

as desired. □

2.6 Sidorenko's conjecture

The famous Sidorenko's conjecture relates the subgraph density to edge density. To state the conjecture, we need some definitions. I thank Joonkyung Lee for teaching me the material in this section.

A *homomorphism* from a graph H to a graph G is a map $f : V(H) \to V(G)$ that preserves adjacency, i.e. if $uv \in E(H)$, then $f(u)f(v) \in E(G)$. Denote by $\mathrm{Hom}(H, G)$ the set of all homomorphisms from H to G and let $\hom(H, G) = |\mathrm{Hom}(H, G)|$. The *homomorphism density* of H in G, or simply H*-density*, is the fraction of maps from $V(H)$ to $V(G)$ that are homomorphisms, that is,

$$t(H, G) = \frac{\hom(H, G)}{|V(G)|^{|V(H)|}}.$$

We call $t(K_2, G)$ the *edge-density* of G.

Conjecture 2.8 (Sidorenko's conjecture)**.** *Let H be a bipartite graph. Then for all graphs G,*

$$t(H, G) \ge t(K_2, G)^{e(H)}. \tag{3}$$

Note that the right-hand-side $t(K_2, G)^{e(H)}$ is the H-density we expect to see in Erdős-Rényi binomial random graphs. Thus, Sidorenko's conjecture states roughly that given edge-density, random graphs minimise H-density, for any bipartite graph H.

Sidorenko's conjecture is still wide open. We will illustrate how to use entropy to prove two cases of this conjecture.

Let us write (3) in a more convenient form. Suppose the host graph G has n vertices. Write $p = t(K_2, G)$ and $v(H) = |V(H)|$. Then $p = t(K_2, G) = \frac{2e(G)}{n^2}$ and (3) becomes

$$\hom(H, G) \ge n^{v(H)} p^{e(H)}.$$

2.6.1 Star of size 2

We start with a toy example of $K_{1,2}$, star of size 2 (or equivalently path of length 2). Let G be an n-vertex graph and $p = t(K_2, G)$ as above. Then Sidorenko's conjecture for $K_{1,2}$ states that

$$\hom(K_{1,2}, G) \ge n^3 p^2. \tag{4}$$

This case follows from a single application of Jensen's inequality. We give a proof using entropy, which is longer. The point is that this entropy argument, though longer, is easier to generalise to handle both when H is a tree or when H is a complete bipartite graph. We encourage the readers to give it a try for the cases (i) $K_{1,b}$, a star of size b, (ii) $H = P_4$, a path on 4 vertices, (iii) $H = K_{2,2} = C_4$, a 4-cycle.

Proof of (4). Sample a random homomorphism (X, Y, Z) from $\mathrm{Hom}(K_{1,2}, G)$ as follows.

- First sample an edge XY uniformly at random. Equivalently, sample a vertex X proportional to its degree and then choose a random neighour Y of X uniformly.
- Then choose a random neighour Z of X uniformly.

The point is that Y and Z are independent conditioning on X, i.e.

$$\mathsf{H}(Y, Z|X) = \mathsf{H}(Y|X) + \mathsf{H}(Z|X),$$

and that both XY and XZ are a random uniform edge. Thus, by maximality,

$$\mathsf{H}(X, Y) = \mathsf{H}(X, Z) = \log\big(\mathrm{hom}(K_2, G)\big) = \log(n^2 p).$$

Then, using maximality, additivity and the conditional independence of Y, Z given X, the entropy of this random homomorphism (X, Y, Z) satisfies

$$\begin{aligned}\log(\mathrm{hom}(K_{1,2}, G)) \geq \mathsf{H}(X, Y, Z) &= \mathsf{H}(Y, Z|X) + \mathsf{H}(X)\\ &= \mathsf{H}(Y|X) + \mathsf{H}(Z|X) + \mathsf{H}(X)\\ &= \mathsf{H}(X, Y) - \mathsf{H}(X) + \mathsf{H}(X, Z) - \mathsf{H}(X) + \mathsf{H}(X)\\ &= \mathsf{H}(X, Y) + \mathsf{H}(X, Z) - \mathsf{H}(X)\\ &\geq 2\log(n^2 p) - \log_2 n\\ &= \log(n^3 p^2),\end{aligned}$$

as desired. □

The key idea here is that Y and Z are made independent conditioning on X and having the same marginal distribution.

2.6.2 Bipartite H with a dominating vertex

In this section, we present an important case of Sidorenko's conjecture due to Conlon, Fox and Sudakov. They proved that Sidorenko's conjecture holds for bipartite H on partite sets A, B if there is a vertex in A adjacent to all vertices in B.

Theorem 2.9. *Let H be a bipartite graph on partite sets A, B with a vertex in A adjacent to all vertices in B. Let G be an n-vertex graph and let $p = t(K_2, G)$. Then*

$$\mathrm{hom}(H, G) \geq n^{v(H)} p^{e(H)}.$$

The original proof uses dependent random choice and tensor power trick. We will present an entropy proof. The following bound on the entropy of a random star will be useful. It can be proved using arguments in the proof of $K_{1,2}$; we leave it as an exercise.

Exercise 2.10. Generate a random star of size b in an n-vertex graph G as follows. First sample a vertex Y proportional to its degree as the center of the star. Then sample $X_1, \ldots, X_b \in N(Y)$ uniformly at random as leaves. Then

$$\mathsf{H}(Y, X_1, \ldots, X_b) \geq \log(n^{b+1} p^b),$$

where $p = t(K_2, G)$.

Proof of Theorem 2.9. Let $z \in A$ be the vertex with $N_{\mathsf{H}}(z) = B$, write $B = [b]$ and let $\mathcal{I} = \{N_{\mathsf{H}}(a) : a \in A \setminus \{z\}\}$ be the collection of sets of neighbours of vertices in $A \setminus \{z\}$. Sample a random homomorphism from $\mathrm{Hom}(H, G)$ as follows.

- First, to embed z, sample a vertex Z proportional to its degree.
- Then, for B, sample vertices $X_1, \dots, X_b \in N_G(Z)$ uniformly at random.
- Finally, to embed $A \setminus \{z\}$, for each $I \in \mathcal{I}$, writing $X_I = (X_i : i \in I)$, sample a vertex $Y_I \in N_G(X_I)$ in such a way that (Y_I, X_I) is distributed as the random star in Exercise 2.10. Equivalently, Y_I is a conditionally independent copy of Z conditioning on X_I.

By maximality, the entropy of this random homomorphism is at most $\log(\hom(H, G))$. It suffices to show that

$$\mathsf{H}\Big(Z, (X_i : i \in [b]), (Y_I : I \in \mathcal{I})\Big) \geq \log\left(n^{v(H)} p^{e(H)}\right).$$

Rewrite the entropy using chain rule:

$$\mathsf{H}\Big(Z, (X_i : i \in [b]), (Y_I : I \in \mathcal{I})\Big) = \mathsf{H}\Big(Z, (X_i : i \in [b])\Big) + \mathsf{H}\Big((Y_I : I \in \mathcal{I})\Big| Z, (X_i : i \in [b])\Big).$$

As $(Z, (X_i : i \in [b]))$ is distributed as in Exercise 2.10, we have for the first term on the right above that

$$\mathsf{H}\Big(Z, (X_i : i \in [b])\Big) \geq \log(n^{b+1} p^b).$$

For the second term, notice that conditioning on $(X_i : i \in [b])$, Z and Y_I, $I \in \mathcal{I}$, are mutually independent. Also, for each $I \in \mathcal{I}$, conditioning on X_I, Y_I and $X_{[b]\setminus I}$ are independent. Thus, by conditional independence and additivity, we have

$$\begin{aligned}
\mathsf{H}\Big((Y_I : I \in \mathcal{I})\Big| Z, (X_i : i \in [b])\Big) &= \mathsf{H}\Big((Y_I : I \in \mathcal{I})\Big| (X_i : i \in [b])\Big) \\
&= \sum_{I \in \mathcal{I}} \mathsf{H}\Big(Y_I \Big| (X_i : i \in [b])\Big) \\
&= \sum_{I \in \mathcal{I}} \mathsf{H}(Y_I | X_I) \\
&= \sum_{I \in \mathcal{I}} \Big(\mathsf{H}(Y_I, X_I) - \mathsf{H}(X_I)\Big).
\end{aligned}$$

Again by Exercise 2.10, $\mathsf{H}(Y_I, X_I) \geq \log(n^{|I|+1} p^{|I|})$. By subadditivity and maximality, $\mathsf{H}(X_I) \leq |I| \log n$. Thus, noting that $|\mathcal{I}| = |A \setminus \{z\}| = v(H) - b - 1$ and that

$$\sum_{I \in \mathcal{I}} |I| = e(H) - d_{\mathsf{H}}(z) = e(H) - b,$$

we get that

$$\begin{aligned}
\mathsf{H}\Big((Y_I : I \in \mathcal{I})\Big| Z, (X_i : i \in [b])\Big) &\geq \sum_{I \in \mathcal{I}} \Big(\log(n^{|I|+1} p^{|I|}) - |I| \log n\Big) \\
&= \sum_{I \in \mathcal{I}} \log(n p^{|I|}) \\
&= \log\left(n^{|\mathcal{I}|} p^{\sum_{I \in \mathcal{I}} |I|}\right) \\
&= \log\left(n^{v(H)-b-1} p^{e(H)-b}\right).
\end{aligned}$$

Finally,

$$\mathsf{H}\Big(Z, (X_i : i \in [b]), (Y_I : I \in \mathcal{I})\Big) \geq \log(n^{b+1} p^b) + \log\left(n^{v(H)-b-1} p^{e(H)-b}\right) = \log(n^{v(H)} p^{e(H)}),$$

as desired. □

2.7 H-colourings

In Sidorenko's conjecture, we are interested in lower bounding the number of homomorphisms from a *fixed* graph H to a *large* graph G. In this section, we will consider the problem of upper bounding the number of homomorphisms from a large graph to a fixed graph instead.

We will study the natural extremal question: given a fixed graph H, which graph, among all those having n vertices and m edges, maximises $\hom(G, H)$? Galvin and Tetali proved the following lovely result for bipartite regular graphs G.

Theorem 2.11. *Let H be a graph with loops allowed but no multiple edges. Then for any n-vertex d-regular bipartite graph G,*

$$\hom(G, H) \leq \hom(K_{d,d}, H)^{\frac{n}{2d}}.$$

This theorem says that for bipartite regular graphs G, $\hom(G, H)$ is maximised when G is a union of $K_{d,d}$. This is a rather general statement; let us see two special cases of it.

Consider the case when $H = K_q$ is a clique, then every homomorphism in $\hom(G, H)$ is a proper q-colouring of G. Thus, a homomorphism to a fixed graph H can be seen as a generalisation of proper colouring. As such, in the literature homomorphisms to a fixed H are also refered as *H-colourings.* Theorem 2.11 then implies the following.

Corollary 2.12. *For any n-vertex d-regular bipartite graph G,*

$$c_q(G) \leq c_q(K_{d,d})^{\frac{n}{2d}},$$

where $c_q(\cdot)$ is the number of proper q-colourings.

When H is the graph on two adjacent vertices u, v with a loop at v, then a homomorphism in $\hom(G, H)$ can be identified with an independent set in G via the preimage of the unlooped vertex u. In this case, Theorem 2.11 implies the following result of Kahn. We will prove Theorem 2.13 below. The same argument works for Theorem 2.11 as well.

Theorem 2.13. *For any n-vertex d-regular bipartite graph G,*

$$i(G) \leq i(K_{d,d})^{\frac{n}{2d}},$$

where $i(\cdot)$ is the number of independent sets.

Proof. Let $V(G) = \mathcal{O} \cup \mathcal{E}$ be a bipartition of G. As G is regular, $|\mathcal{O}| = |\mathcal{E}| = \frac{n}{2}$. Write $\mathcal{I}(G)$ for the set of all independent sets in G. Let $X \sim \mathcal{I}(G)$ be a uniform random independent set drawn from G and let $Y \sim \mathcal{I}(K_{d,d})$. By maximality, it suffices to show that

$$\log i(G) = \mathsf{H}(X) \leq \frac{n}{2d}\mathsf{H}(Y) = \log i(K_{d,d})^{\frac{n}{2d}}.$$

Writing X_v for the indicator function of $\{v \in X\}$, we can view X has a random vector $(X_{\mathcal{O}}, X_{\mathcal{E}})$, where $X_{\mathcal{O}} = (X_v : v \in \mathcal{O})$ and $X_{\mathcal{E}} = (X_v : v \in \mathcal{E})$. Then by chain rule,

$$\mathsf{H}(X) = \mathsf{H}(X_{\mathcal{E}}) + \mathsf{H}(X_{\mathcal{O}}|X_{\mathcal{E}}).$$

We bound the first term $\mathsf{H}(X_{\mathcal{E}})$ using Shearer's lemma. For this, we need to take a random set F of coordinates in $\mathcal{E}$. Let $Z \sim \mathcal{O}$ and let $F = N(Z)$, then $\mathsf{Pr}(v \in F) = \mu = \frac{d}{n/2}$ for each $v \in \mathcal{E}$. Thus, Shearer's lemma infers

$$\mathsf{H}(X_{\mathcal{E}}) \leq \frac{n/2}{d}\mathsf{E}_F\mathsf{H}(X_F) = \frac{n/2}{d}\mathsf{E}_Z\mathsf{H}(X_{N(Z)}) = \frac{1}{d}\sum_{v \in \mathcal{O}} \mathsf{H}(X_{N(v)}).$$

The second term can be bounded using subadditivity and dropping conditioning:[2]

$$\begin{aligned}\mathsf{H}(X_{\mathcal{O}}|X_{\mathcal{E}}) &\le \sum_{v\in\mathcal{O}} \mathsf{H}(X_v|X_{\mathcal{E}})\\ &\le \sum_{v\in\mathcal{O}} \mathsf{H}(X_v|X_{N(v)}).\end{aligned}$$

Thus, we have

$$\mathsf{H}(X) \le \frac{1}{d}\sum_{v\in\mathcal{O}} \Big(\mathsf{H}(X_{N(v)}) + d\cdot \mathsf{H}(X_v|X_{N(v)})\Big).$$

It then suffices to show that for any $v \in \mathcal{O}$,

$$\mathsf{H}(X_{N(v)}) + d\cdot \mathsf{H}(X_v|X_{N(v)}) \le \mathsf{H}(Y).$$

The above inequality follows from maximality. To see this, note that X induces a random independent set W in $K_{d,d}$ as follows. Write A, B for the two partite sets of $K_{d,d}$. Let W be drawn from $\mathcal{I}(K_{A,B})$ in such a way that

- its marginal distribution on B is the same as $X_{N(v)}$, i.e. W_B and $X_{N(v)}$ are identically distributed; and
- each vertex $u \in A$ has the same conditional marginal distribution as $X_v|X_{N(v)}$, i.e. $W_u|W_B$ and $X_v|X_{N(v)}$ are identically distributed.

By the choice of W, invariance and maximality, we get

$$\mathsf{H}(X_{N(v)}) + d\cdot \mathsf{H}(X_v|X_{N(v)}) = \mathsf{H}(W) \le \mathsf{H}(Y),$$

as desired. □

2.8 Counting matroids

A matroid is a structure that abstracts and generalises the notion of linear independence in vector spaces. Formally, a *matroid* is a pair $(E, \mathcal{B})$, where E is a finite set and $\mathcal{B} \subseteq 2^E$ is a nonempty collection of subsets of E, satisfying the following axiom:

- *Base exchange.* For any $B, B' \in \mathcal{B}$ and any $e \in B \setminus B'$, there exists $f \in B' \setminus B$ such that $B \setminus \{e\} \cup \{f\} \in \mathcal{B}$.

We call E the *ground set*, elements in $\mathcal{B}$ the *bases*, and subsets of bases are *independent sets*. One can define a matroid as a family of independent sets that are downward closed. A subset of the ground set E is called *dependent set* if it is not independent. Base exchange implies that all bases in a matroid have the same cardinality, which we call the *rank* of the matroid.

Denote by $m_{n,r}$ the number of matroids of rank r on the ground set $[n]$ and by m_n the number of all matroids on $[n]$. Clearly, $m_n \le 2^{2^n}$, or $\log\log m_n \le n$. Knuth gave a construction showing that $m_n \ge 2^{\frac{1}{2}\binom{n}{n/2}}$, or

$$\log\log m_n \ge n - \frac{3}{2}\log n - O(1).$$

We will prove the following upper bound due to Bansal, Pendavingh and van der Pol, which has the same first two leading terms as in Knuth's lower bound.

[2] Note that the second inequality below is in fact an equality. Indeed, as X is chosen uniformly, conditioning on $X_{N(v)}$, X_v has nothing to do with other vertices $\mathcal{E} \setminus N(v)$. This is the *spatial Markov property* that for choosing independent sets, the state of v is only determined by the boundary condition on $N(v)$. This key property drives not only the entropy argument here but also an argument using hard-core model in statistical physics that we shall see later.

Theorem 2.14. *The number of matroids on $[n]$ is at most*

$$\log\log m_n \leq n - \frac{3}{2}\log n + \log\log n + O(1)$$

The strategy is to use entropy method (Shearer's lemma) to bound the number of matroids of higher ranks by the number of matroids of rank 2. Let us then first bound $m_{n,2}$. To warm up, note that $m_{n,0} = 1$ for any n, as the only matroid of rank 0 is $\mathcal{B} = \{\varnothing\}$; and $m_{n,1} = 2^n - 1$, as any nonempty $\mathcal{B} \subseteq \binom{[n]}{1}$ satisfies base exchange. For $m_{n,2}$, we only need the following crude upper bound:

$$m_{n,2} \leq n^n. \tag{5}$$

To see this bound, we need the following lemma, stating that in a rank-2 matroid, for pairs of elements in the ground set, being dependent is an equivalence relation.

Lemma 2.15. *Let $(E, \mathcal{B})$ be a rank-2 matroid, then there is a partition $E_0, E_1, \ldots, E_k$ of E such that*

$$\mathcal{B} = \Big\{\{e, e'\} : e \in E_i, e' \in E_j, ij \in \binom{[k]}{2}\Big\}.$$

Proof. Let E_0 be the set of all singletons that are dependent sets, that is,

$$E_0 = \{e \in E : e \notin B \text{ for any } B\} = E \setminus \big(\cup_{B\in\mathcal{B}} B\big).$$

It suffices to show that being dependent is an equivalence relation for pairs and let E_is being the equivalent classes, i.e. for any $e, f, g \in E \setminus E_0$, if $ef, eg \notin \mathcal{B}$, then $fg \notin \mathcal{B}$. Suppose $ef, eg \notin \mathcal{B}$, but $fg \in \mathcal{B}$. As $e \notin E_0$, there exists some $h \notin \{e, f, g\}$ such that $eh \in \mathcal{B}$. Then base exchange fails for $B = eh$, $B' = fg$ and $h \in B \setminus B'$. □

Lemma 2.15 provides an injective map from the set of rank-2 matroids to partitions of E. Thus, $m_{n,2}$ is at most the number of partitions of $[n]$, proving (5).

Here is the reduction step to rank-2 matroids.

Lemma 2.16. *Let $0 \leq t \leq r \leq n$, then*

$$\frac{1}{\binom{n}{r}}\log(m_{n,r} + 1) \leq \frac{1}{\binom{n-t}{r-t}}\log(m_{n-t,r-t} + 1).$$

In particular, taking $t = r - 2$, we have

$$\frac{1}{\binom{n}{r}}\log(m_{n,r} + 1) \leq \frac{1}{\binom{n-r+2}{2}}\log(m_{n-r+2,2} + 1).$$

Before proving Lemma 2.16, let us see how it implies Theorem 2.14.

Proof of Theorem 2.14. By Lemma 2.16 and (5), we get that

$$\log(m_{n,r} + 1) \leq \frac{\log(m_{n-r+2,2} + 1)}{\binom{n-r+2}{2}}\binom{n}{r} \leq \frac{(n+1)\log(n+1)}{\binom{n-r+2}{2}}\binom{n}{r} = \frac{2\log(n+1)}{n+2}\binom{n+2}{r}.$$

As $m_n = \sum_{r=0}^{n} m_{n,r} \leq (n+1)\max_r m_{n,r}$, we see that

$$\begin{aligned}\log m_n &\leq \log(n+1) + \max_r \log m_{n,r}\\ &\leq \log(n+1) + \frac{2\log(n+1)}{n+2}\binom{n+2}{\lfloor (n+2)/2 \rfloor} = O\Big(2^n n^{-3/2}\log n\Big).\end{aligned}$$

Thus, $\log\log m_n \leq n - \frac{3}{2}\log n + \log\log n + O(1)$ as desired. □

For the proof of Lemma 2.16, we need to introduce contractions on matroids. First a piece of notation: for a set E, define

$$\mathcal{M}_{E,r} = \left\{\mathcal{B} \subseteq \binom{E}{r} : \mathcal{B} \text{ satisfies base exchange.}\right\}$$

Note that if $|E| = n$, then $|\mathcal{M}_{E,r}| = m_{n,r} + 1$, as $\mathcal{M}_{E,r}$ contains the empty set apart from all the rank-r matroids.

Let $M = (E, \mathcal{B})$ be a matroid. If $T \subseteq E$ is contained in some basis of M, then *contracting* T yields another matroid $M/T = (E \setminus T, \mathcal{B}/T)$, where

$$\mathcal{B}/T = \{B \setminus T : B \in \mathcal{B}, T \subseteq B\}.$$

If T is not contained in any basis of $M = (E, \mathcal{B})$, then $\mathcal{B}/T = \varnothing$. Thus, for any $\mathcal{B} \in \mathcal{M}_{E,r}$ and any $T \subseteq E$,

$$\mathcal{B}/T \in \mathcal{M}_{E\setminus T, r-|T|}.$$

Proof of Lemma 2.16. Let $E = [n]$ and draw $X \sim \mathcal{M}_{E,r}$. Thus, by maximality, $\mathsf{H}(X) = \log(m_{n,r} + 1)$. We can view X as the random vector $(X_R : R \in \binom{E}{r})$, where X_R is the indicator function for $R \in X$.

To apply Shearer's lemma, the projections we shall do come from contractions of X. For each $T \subseteq E$ of size t, the contraction X/T takes values in $\mathcal{M}_{E\setminus T, r-t}$ and so by maximality,

$$\mathsf{H}(X/T) \leq \log(m_{n-t,r-t} + 1).$$

Note that X/T is the projection of X to the set of coordinates $F(T) := \{R \in \binom{E}{r} : T \subseteq R\}$, i.e.

$$X/T = X_{F(T)}.$$

Now, let $T \sim \binom{E}{t}$ be a uniform random t-set and $F(T)$ be the induced set of random coordinates. Then for each $R \in \binom{E}{r}$,

$$\mathsf{Pr}(R \in F(T)) = \mathsf{Pr}(T \subseteq R) = \mu = \frac{\binom{r}{t}}{\binom{n}{t}} = \frac{\binom{n-t}{r-t}}{\binom{n}{r}}.$$

Thus, Shearer's lemma implies that

$$\begin{aligned}
\log(m_{n,r} + 1) = \mathsf{H}(X) &= \mathsf{H}\Big(X_R : R \in \binom{E}{r}\Big) \\
&\leq \frac{\binom{n}{r}}{\binom{n-t}{r-t}} \mathsf{E}_{F(T)} \mathsf{H}(X_{F(T)}) \\
&= \frac{\binom{n}{r}}{\binom{n-t}{r-t}} \mathsf{E}_T \mathsf{H}(X/T) \\
&\leq \frac{\binom{n}{r}}{\binom{n-t}{r-t}} \log(m_{n-t,r-t} + 1),
\end{aligned}$$

as desired. □

2.9 Sunflower

A P-*sunflower* is a family of P sets with identical pairwise intersection, that is, $\{S_1, \ldots, S_P\}$ such that for any $ij \in \binom{[P]}{2}$, $S_i \cap S_j = A$ for some *core* set A. We call the sets in a sunflower *petals*. Note that A is allowed to be empty, i.e. P pairwise disjoint sets also form a P-sunflower. Erdős and Rado made the following well-known conjecture on how large a family of k-sets (k-uniform hypergraphs, or simply k-graphs) can be without containing a sunflower.

Conjecture 2.17 (Sunflower conjecture)**.** *Let $P \geq 3$. There exists $C = C(P)$ such that any k-uniform hypergraph $\mathcal{F}$ without a P-sunflower has at most C^k edges.*

Erdős and Rado proved an upper bound of $(P-1)^k \cdot k!$. A recent breakthrough of Alweiss, Lovett, Wu and Zhang gave an upper bound of $\big(\log k \cdot (P \log\log k)^{O(1)}\big)^k$. Rao later provided a simpler proof using Shannon's noiseless coding theorem and obtained a slightly better upper bound of $(O(1) \cdot P \log(Pk))^k$. More recently, Frankston, Kahn, Narayanan, and Park builds on these developments to resolve a conjecture of Talagrand on expectation threshold in probabilistic combinatorics. Tao gave a proof of Rao's bound using entropy. The advantage of the entropy argument is that entropy smooths things out and one can prove the statement in one go without having to distinguish typical/atypical cases as in Rao or Frankston, Kahn, Narayanan, and Park's proofs.

If we take $P = O(1)$, then the above bound is $O(\log k)^k$. It remains open even for 3-sunflower, whether we can get down to $O(1)^k$.

We will present the following slightly weaker bound using Tao's entropy proof.

Theorem 2.18. *Any k-uniform hypergraph with more than $(P \log(Pk))^{O(k)}$ edges contains a P-sunflower.*

Throughout the rest of this section, given a set X, we write $X_\delta = \mathrm{Binom}(X, \delta)$ for the random binomial subset in which each element in X is included with probability δ independent of others. Recall that given a set A, we write $W \sim A$ for a random variable W uniformly sampled from A. We will abuse the notation slightly and also write $W \sim Z$ to mean two random variables W and Z having the same distribution.

We will later pass from binomial random model to the easier-to-work uniform random model. For this, we need the following definition. An *empirical sequence* $X_1, X_2, \dots,$ for a random variable X is a sequence taking the values from the same set as X with empirical samples converging in distribution to X, that is, let n be drawn uniformly from $[N]$, then

$$\Pr(X_n = x) = \Pr(X = x) + o_N(1).$$

2.9.1 Reduction to spread hypergraphs

We first show that, via induction, we may consider only hypergraphs with certain pseudorandomness property. Below is the formal definition. In other words, spread condition impose upper bounds on codegree of sets in the hypergraph.

Definition 2.19 (R-spread)**.** Let $R > 1$. A k-uniform hypergraph $\mathcal{H}$ is *R-spread* if for any $S \neq \varnothing$,

$$d(S) = \big|\{T \in \mathcal{H} : T \supseteq S\}\big| \leq \frac{e(\mathcal{H})}{R^{|S|}}.$$

Let $\mathcal{H}$ be a k-graph with no P-sunflower. Say we want to show an upper bound $e(\mathcal{H}) \leq R^k$. If $e(\mathcal{H}) > R^k$ and $\mathcal{H}$ is not R-spread, then by definition, there exists some nonempty S lying in more than $\frac{e(\mathcal{H})}{R^{|S|}} > R^{k-|S|}$ edges, say $\{T_i\}_{i \in I}$, of $\mathcal{H}$. Then the link of S, that is, the $(k-|S|)$-graph $\mathcal{H}/S$ with edge set $\{T_i \setminus S\}_{i \in I}$, has more than $R^{k-|S|}$ edges and we can induct on the uniformity k to get a sunflower in $\mathcal{H}/S$, which corresponds to a sunflower in $\mathcal{H}$ with core containing S.

Thus, to prove Theorem 2.18, it suffices to show the following.

Theorem 2.20. *There exists $C > 0$ such that every k-uniform R-spread hypergraph $\mathcal{H}$ with $R = (P \log(Pk))^C$ and $e(\mathcal{H}) > R^k$ contains a P-sunflower.*

We shall find in such spread hypergraph a P-sunflower with empty core, that is, P pairwise disjoint sets. We will phrase things in a probabilistic way.

Definition 2.21 (R-spread, probabilistic version)**.** Let $R > 1$. A random set T is *R-spread*, , if

$$\Pr(S \subseteq T) \le R^{-|S|}, \quad \text{for any set } S.$$

A hypergraph is R-spread if the random edge drawn uniformly from its edge set is R-spread.

The following is the key lemma.

Lemma 2.22. *Let $R > 1$ and $0 < \delta < 1$. Let A be a random R-spread subset of X, and $V \sim \mathrm{Binom}(X, \delta)$ be a random subset of X, independent of A. Then there exists another random subset A' of X such that*

- *A' has the same distribution as A;*
- *$A' \setminus V \subseteq A$;*
- $\mathsf{E}|A' \setminus V| \le \frac{4+\log\frac{1}{\delta}}{\log R} \cdot \mathsf{E}|A|$.

The interesting thing here is that as V and A are independent, we only have $\mathsf{E}|A \setminus V| = (1-\delta)\mathsf{E}|A|$, but by taking some $A' \sim A$, we can replace the $1 - o(1)$ factor to $o(1)$ factor, provided that $R \gg 1/\delta$.

Let us first sketch how Lemma 2.22 implies Theorem 2.20.

Proof sketch of Theorem 2.20. Take an R-spread hypergraph $\mathcal{H}$ as in Theorem 2.20 and let $X = V(\mathcal{H})$ and $A \sim E(\mathcal{H})$, then A is a random R-spread subset of X and $|A| = k$. Set $\delta \approx \frac{1}{P\log(Pk)}$ and $R = (P\log(Pk))^C$ for large $C > 0$ so that $\frac{4+\log\frac{1}{\delta}}{\log R} \le \frac{1}{2}$. Iterating Lemma 2.22 $m \approx \log(Pk)$ rounds, we get that if $V \sim \mathrm{Binom}(X, \frac{1}{P})$, then for some $A' \sim A$,

$$\mathsf{Pr}\big(|A' \setminus V| \ge 1\big) \le \mathsf{E}|A' \setminus V| < \frac{1}{P}. \tag{6}$$

Now randomly partition X into $V_1 \cup \ldots \cup V_P$ by placing every element of X into a uniformly chosen V_i, $i \in [P]$, independent of others. Then for each $i \in [P]$, $V_i \sim \mathrm{Binom}(X, \frac{1}{P})$. Thus, by (6) and union bound, with positive probability, we can find P pairwise disjoint edges, one in each V_i, yielding a desired P-sunflower. □

2.9.2 Incorporate spreadness in entropy

To prove Lemma 2.22, we need a bit more on the theory of entropy.

Conditional entropy of a random subset. Let X, Y be random variables. Recall that by maximality, if X takes values in a set S_Y that depends on Y, then $\mathsf{H}(X|Y) \le \mathsf{E}\log|S_Y|$. Using this, we get the following upper bound on the conditional entropy of a random subset:

$$\mathsf{H}(B|A) \le \mathsf{E}|A|, \quad \text{for any random subset } B \text{ of a random set } A. \tag{7}$$

Mutual information. The *mutual information* of two random variables X, Y is

$$\begin{aligned} \mathsf{I}(X;Y) &= \mathsf{H}(X) - \mathsf{H}(X|Y) = \mathsf{H}(Y) - \mathsf{H}(Y|X) \\ &= \mathsf{H}(X) + \mathsf{H}(Y) - \mathsf{H}(X,Y) \\ &= \mathsf{H}(X,Y) - \mathsf{H}(X|Y) - \mathsf{H}(Y|X). \end{aligned}$$

As suggested by its name, $\mathsf{I}(X;Y)$ measures the information shared by X and Y, which is always non-negative by subadditivity. If Y is determined by X, i.e. $Y = f(X)$, then $\mathsf{I}(X;Y) = \mathsf{H}(Y)$; while if X, Y are independent, then $\mathsf{I}(X;Y) = 0$.

Let us first see how much mutual information a random set has with its random subset. Consider a random set A drawn uniformly from $[n]$ and a random subset $B \subseteq A$. Then the mutual information B has with A is $\mathsf{E}|B|$:

$$\begin{aligned}\mathsf{I}(A;B) &= \mathsf{H}(A) - \mathsf{H}(A|B)\\ &= \mathsf{E}_b\Big(\mathsf{H}(A) - \mathsf{H}(A||B| = b)\Big)\\ &= \mathsf{E}_b\big(n - (n-b)\big)\\ &= \mathsf{E}|B|.\end{aligned}$$

The following lemma states that, compared to the usual case above, a spread set has *lots* of mutual information (additional $\log R$ factor) with its large random subsets. Combinatorially, if B is large, then there are fewer choices to extend it to a spread A, meaning that the marginal information $\mathsf{H}(A|B)$ is small and so $\mathsf{I}(A;B) = \mathsf{H}(A) - \mathsf{H}(A|B)$ is large.

Lemma 2.23. *Let A be a random R-spread set with $R > 1$. If A is uniformly chosen, then for any random subset $B \subseteq A$, we have*

$$\mathsf{I}(A;B) = \mathsf{H}(A) - \mathsf{H}(A|B) \geq \log R \cdot \mathsf{E}|B|.$$

By the above lemma, to show, for certain random subset B of a spread A, that $\mathsf{E}|B|$ is small compared to $\mathsf{E}|A|$, we can try to upper bound the mutual information $\mathsf{I}(A;B)$.

As we remark before, spread is a pseudorandom property, and genuin random sets possess this property. In particular, random sets with density δ is $(1/\delta)$-spread.

Lemma 2.24. *Let $0 < \delta < 1$ and W be uniformly chosen from $\binom{X}{\delta|X|}$. Then the following holds.*

- *For any random set $B \subseteq W$,*

$$\mathsf{I}(W,B) = \mathsf{H}(W) - \mathsf{H}(W|B) \geq \log\frac{1}{\delta} \cdot \mathsf{E}|B|.$$

- *(Aborption) For any random set $B \subseteq X$,*

$$\mathsf{H}(W \cup B) \leq \mathsf{H}(W) + 1 + \left(1 + \log\frac{1}{\delta}\right)\mathsf{E}|B|. \tag{8}$$

Relative product. A key idea in the proof is to utilise *relative product* (X, X'), in which X' is a conditionally independent copy of X subject to certain constraint $f(X) = f(X')$ for a given function f. Using relative product, we can rewrite the entropy of X as follows:

$$\begin{aligned}\mathsf{H}(X) &= \mathsf{H}(X, f(X))\\ &= \mathsf{H}(X|f(X)) + \mathsf{H}(f(X))\\ &= \mathsf{H}(X|X') + \mathsf{H}(f(X)),\end{aligned} \tag{9}$$

where the first and second equalities follow from invariance and additivity respectively, and the last equality is due to the conditional independence of X and X' (as in the only relevant information X' can provide for X is $f(X') = f(X)$).

2.9.3 Conditionally independent copy

Proof of Lemma 2.22. If A is empty, then we can take $A' = A$. So we can condition on the event that A is non-empty, and assume that $\mathsf{E}|A| \geq 1$. View $X = [|X|]$ and take large $N_1, N_2 \gg |X|$.

We first pass from the binomial model $V \sim \mathrm{Binom}(X, \delta)$ to the uniform model $W \sim \binom{[N_2]}{\delta N_2}$. For this, take an empirical sequence $A_1, A_2, \dots$ for A with $A_i \subseteq [|X|]$. Let $n \sim [N_1]$, so

$$\mathsf{E}|A_n| \geq \mathsf{E}|A| - o(1) \geq 1 - o(1). \tag{10}$$

Then take $W \sim \binom{[N_2]}{\delta N_2}$ independent of n. Note that $W \cap [|X|]$ converges in distribution to V. Thus, it suffices to find $n' \sim [N_1]$ such that, as $N_1, N_2 \to \infty$,

$$\mathsf{E}|A_{n'} \setminus W| \leq \frac{4 + \log \frac{1}{\delta}}{\log R} \cdot \mathsf{E}|A_n| + o(1).$$

To get this, we (cleverly) take a conditionally independent copy (n', W') of (n, W) subject to having the same union

$$A_n \cup W = A_{n'} \cup W' \qquad \Longrightarrow \qquad A_{n'} \setminus W \subseteq A_n \cap A_{n'}.$$

Then it suffices to show

$$\mathsf{E}|A_n \cap A_{n'}| \leq \frac{4 + \log \frac{1}{\delta}}{\log R} \cdot \mathsf{E}|A_n| + o(1).$$

By Lemma 2.23 with $(A, B)_{2.23} = (A_n, A_n \cap A_{n'})$, it amounts to proving

$$\mathsf{I}(A_n; A_n \cap A_{n'}) = \mathsf{H}(n) - \mathsf{H}(n|A_n \cap A_{n'}) \leq \Big(4 + \log \frac{1}{\delta}\Big)\mathsf{E}|A_n| + o(1). \tag{11}$$

Let us start with upper bounding $\mathsf{H}(n)$ via bounding $\mathsf{H}(n, W)$. We do so with the help of the conditionally independent copy (n', W') via (9) with $((n, W), (n', W'), A_n \cup W)$ playing the role of $(X, X', f(X))$:

$$\begin{aligned}\mathsf{H}(n, W) &= \mathsf{H}(n, W|n', W') + \mathsf{H}(A_n \cup W) \\ &\leq \mathsf{H}(n, W|n', W') + \mathsf{H}(W) + \Big(2 + \log \frac{1}{\delta}\Big)\mathsf{E}|A_n| + o(1),\end{aligned}$$

where the last inequality follows from (8) and (10). By the independence of n and W, $\mathsf{H}(n, W) = \mathsf{H}(n) + \mathsf{H}(W)$. Thus we get

$$\mathsf{H}(n) \leq \mathsf{H}(n, W|n', W') + \Big(2 + \log \frac{1}{\delta}\Big)\mathsf{E}|A_n| + o(1). \tag{12}$$

We are left to bound $\mathsf{H}(n, W|n', W')$ and relate it to the subset $A_n \cap A_{n'}$. We bound $\mathsf{H}(n, W|n', W')$ by first choosing $A_n \cap A_{n'}$, then n, then $A_n \setminus W$, which determines W:

$$\mathsf{H}(n, W|n', W') \leq \mathsf{H}(A_n \cap A_{n'}|n', W') + \mathsf{H}(n|A_n \cap A_{n'}, n', W') + \mathsf{H}(W|n, A_n \cap A_{n'}, n', W').$$

We can bound the 1st and 3rd terms using (7) and dropping conditioning:

$$\mathsf{H}(A_n \cap A_{n'}|n', W') \leq \mathsf{H}(A_n \cap A_{n'}|n') \leq \mathsf{E}|A_n|,$$

and

$$\begin{aligned}\mathsf{H}(W|n, A_n \cap A_{n'}, n', W') &\leq \mathsf{H}(W|n, n', W') \\ &= \mathsf{H}(A_n \setminus W|n, n', W') \\ &\leq \mathsf{H}(A_n \setminus W|n) \\ &\leq \mathsf{E}|A_n|,\end{aligned}$$

where the equality holds as, given A_n and the union $A_n \cup W$ (from n', W'), W determines $A_n \setminus W$ and vice versa. Now replacing the 2nd term with the larger one $\mathsf{H}(n|A_n \cap A_{n'})$, we arrive at

$$\mathsf{H}(n, W|n', W') \leq \mathsf{H}(n|A_n \cap A_{n'}) + 2\mathsf{E}|A_n|.$$

Plus this back in (12), we get the desired bound (11). □